Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora

Sylwia OZDOWSKA
Equipe de Recherche en Syntaxe et Sémantique
Université Toulouse le Mirail
5 allées Antonio Machado
31058 Toulouse Cedex 1 France
ozdowska@univ-tlse2.fr

Abstract

We present a word alignment procedure based on a syntactic dependency analysis of French/English parallel corpora called "alignment by syntactic propagation". Both corpora are analysed with a deep and robust parser. Starting with an anchor pair consisting of two words which are potential translations of one another within aligned sentences, the alignment link is propagated to the syntactically connected words. The method was tested on two corpora and achieved a precision of 94.3 and 93.1% as well as a recall of 58 and 56%, respectively for each corpus.

1 Introduction

It is now an aknowledged fact that parallel corpora, i.e. corpora made of texts in one language and their translation in another language, are well suited in particular to cope with the problem of the construction of bilingual resources such as bilingual lexicons or terminologies. Several works have focused on the alignment of units which are smaller than a sentence, for instance words or phrases, as to produce bilingual word, phrase or term associations. A common assumption is that the alignment of words or phrases raises a real challenge, since it is "neither one-to-one, nor sequential, nor compact", and thus "the correspondences are fuzzy and contextual" (Debili, 1997). Indeed, it is even often diffult for a human to determine which source unit correspond to which target unit within aligned sentences (Och and Ney, 2003).

Most alignment systems working on parallel corpora rely on statistical models, in particular the EM ones (Brown, Della Pietra and Mercer, 1993). Quite recently attempts have been made in order to incorporate different types of linguistic information sources into word and phrase alignment systems. The idea is to take into account the specific problems arising from the alignment at the word or phrase level mentioned in particular by Debili (1997). Different types of linguistic knowledge are exploited: morphological, lexical and syntactic ones. In the method described in this article, the syntactic information is the kernel of the alignment process. Indeed, syntactic relations identified on both sides of the French/English parallel corpus with a deep and robust parser are used to find out new correspondences between words or to confirm existing ones in order to achieve a high accuracy alignment. We call this procedure "alignment by syntactic propagation".

2 State of the art

2.1 Term alignment

Two kinds of methods have been basically proposed in order to address the problem of bilingual lexicon extraction. On the one hand, terms are recognized in both source and target language and then they are mapped to each other (Daille, Gaussier and Langé, 1994). On the other hand, only source terms are extracted and the target ones are discovered through the alignment process (Gaussier, 1998; Hull, 2001). The alignment between terms is obtained either by computing association probabilities (Gaussier, 1998 ; Daille, Gaussier and Langé, 1994) or by identifying, for a given source term, a sequence of words in the target language which is likely to contain or to correspond to its translation (Hull, 2001). In so far as the precision rate may be affected by the number of alignments obtained (Daille, Gaussier and Langé, 1994; Gaussier, 1998), the results achieved basically range between 80% and 90%, for the first 500 alignments. As for the method described in (Hull, 2001), the precison reported is 56%.

It should be noticed that the use of linguistic knowledge is most of the time restricted to the term recognition stage. This kind of knowledge is quite rarely taken into account within the very alignment process, except for the approach implemented by Daille, Gaussier and Langé (1994), which try to take advantage of

correspondences between the syntactic patterns defined for each language.

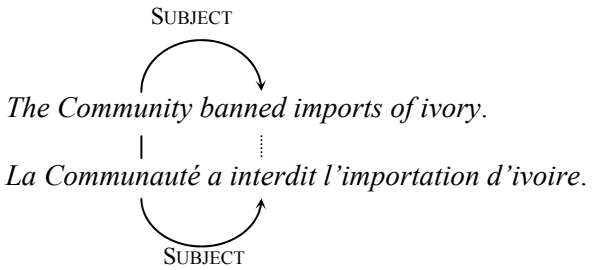
2.2 Word alignment

Quite recently attempts have been made in order to incorporate different types of linguistic information sources into word alignment systems and to combine them with statistical knowledge. Various and more or less complex sources of linguistic knowledge are exploited: morphological, lexical (Arhenberg, Andersson and Merkel, 2000) and syntactic knowledge (Wu, 2000; Lin and Cherry, 2003). The contribution of these information sources to the alignment process with respect to the statistical data varies according to the considered system. However, as pointed out by Arhenberg, Andersson and Merkel (2000) as well as Lin and Cherry (2003), the introduction of linguistic knowledge leads to a significant improvement in alignment quality. In the first case, the accuracy goes from 91% for a baseline configuration up to 96.7% for a linguistic knowledge based one. In the second, the precision rate is increased from 82.7% up to 89.2% and the improvement noticed have been confirmed within the framework of an evaluation task (Mihalcea and Pedersen, 20003).

For our part, we propose a method in which the syntactic information plays a major role in the alignment process, since syntactic relations are used to find out new correspondences between words or to confirm the existent ones. We chose this approach in order to achieve a high accuracy alignment both at word and phrase level. Indeed, we aim at capturing frequent alignments between words and phrases as well as those involving sparse or corpus specific ones. Moreover, as stressed in previous works, using syntactic dependencies seems to be particularly well suited to solve *n*-to-1 or *n*-to-*m* alignments (Fluhr, Bisson and Elkateb, 2000) and to cope with the problem of linguistic variation and non correspondence across languages, for instance when aligning terms (Gaussier, 2001).

3 Starting hypothesis

We take as a starting point the hypothesis formulated by Debili and Zribi (1996) according to which "*paradigmatic connections can help to determine syntagmatic relations, and conversely*"[1]. More precisely, the idea is that one can make use of syntactic relations to validate or invalidate the existence of alignment links, on the one hand, and to create new ones, on the other hand. The reasoning is as follows : if there is a pair of anchor words, i.e. if two words $w1_i$ (*community* in the example) and $w2_m$ (*communauté*) are aligned at the sentence level, and if there is a syntactic relation standing between $w1_i$ (*community*) and $w1_j$ (*ban*) on the one hand, and between $w2_m$ (*communauté*) and $w2_n$ (*interdire*) on the other hand, then the alignment link is propagated from the anchor pair (*community*, *communauté*) to the words (*ban*, *interdire*). We call this procedure "alignment by syntactic propagation".



In the rest of this article, we describe the overall design and implementation of the syntactic propagation process and the results of applying it to two parsed French/English parallel corpora: INRA and JOC.

4 Corpus processing

The alignment by syntactic propagation was tested on two different parallel corpora aligned at the sentence level: INRA and JOC. The first corpus was constituted at the National Institute for Agricultural Research (INRA)[2] to enrich a bilingual terminology database exploited by translators. It comprises about 300,000 words and mainly consists of research and popular-science papers, press releases.

The JOC corpus was provided by the ARCADE project, a campaign devoted to the evaluation of parallel text alignment systems (Veronis and Langlais, 2000). It contains written questions on a wide variety of topics addressed by members of the European Parliament to the European Commission and corresponding answers published by the Official Journal of the European Community in nine official languages. A portion of about 400,000 words of the French and English parts were used in the framework of the ARCADE evaluation task.

The corpus processing was carried out by a French/English parser: SYNTEX (Bourigault and Fabre, 2000; Frérot, Fabre and Bourigault, 2003). SYNTEX is a dependency parser whose input is a

[1]Our translation of the French version *« les liaisons paradigmatiques peuvent aider à déterminer les relations syntagmatiques, et inversement »*.

[2] We are grateful to A. Lacombe who allowed us to use this corpus for research purposes.

POS tagged[3] corpus–meaning each word in the corpus is assigned a lemma and grammatical tag. The parser identifies syntactic dependencies in the sentences of a given corpus, for instance subjects, direct and indirect objects of verbs. Once all syntactic dependencies have been identified, a set of words and phrases is extracted out of the corpus.

Both versions of the parser–the French one and the English one–are being developed according to the same procedures and architecture. The parsing is performed independently in each language, yet the outputs are quite homogeneous since the syntactic dependencies are identified and represented in the same way in both languages. In this respect, the alignment method proposed is different from the ones developed by Wu (2000) as well as Lin and Cherry (2003): the former is based on synchronous parsing while the letter uses a dependency tree generated only in the source language.

In addition to parsed French/English corpus aligned at the sentence level, the syntactic alignment requires pairs of anchor words be identified prior to propagation as to start the process. In this study, we chose to extract a lexicon out of the corpus, the anchor pairs being located both by projecting the lexicon at the level of aligned sentences and processing the identical and fuzzy cognates.

5 Identification of anchor pairs

To derive a list of words which are likely to be used to initiate the syntactic propagation process out of the corpus, we implemented a widely used method described notably in (Gale and Church, 1991; Ahrenberg, Andersson and Merkel, 2000) which is based on the assumption that the words which appear frequently in aligned text segments are potential translation equivalents. For each source (English) and target (French) unit, respectively u_1 and u_2, extracted by SYNTEX, the translation equivalents are searched for by counting co-occurrences of (u_1, u_2) in aligned sentences in comparison with their overall occurrences in the corpus and then an association score is computed. In this study, we chose the Jaccard association score which is calculated as follows:

$$j(u_1, u_2) = \frac{f(u_1, u_2)}{f(u_1) + f(u_2) - f(u_1, u_2)}$$

The association score is computed provided the number of overall occurrences of u1 and u2 is higher than 4 since statistical techniques have proved to be particularly efficient when aligning frequent units. Moreover, the alignments are filtered according to the $j(u_1, u_2)$ value, provided the latter is higher than 0.2. Then, two tests, based on cognate recognition and mutual correspondence condition (Altenberg, 1999), are applied as to filter spurious associations out of the initial lexicon.

The identification of anchor pairs, consisting of words which are translation equivalents within aligned sentences, combines both the projection of the initial lexicon and the recognition of cognates for words which have not been taken into account in the lexicon. These pairs are used as the starting point of the propagation process.

Table 1 gives some characteristics of the two corpora as for the number of aligned sentences, the overall number of anchor pairs identified, the average number of anchor pairs per sentence pair as well as the precision rate[4] of the anchor pairs. It can be seen that a high number of anchor pairs has been identified per sentence for both corpora with a high accuracy.

	INRA	JOC
aligned sentences	7056	8774
anchor pairs	42570	58771
words/source sentence	21	25
words/target sentence	24	30
anchor pairs/sentence	6.38	6.77
precision (%)	98	99.3

Table 1: The identification of anchor pairs

6 Syntactic propagation

6.1 Two types of propagation

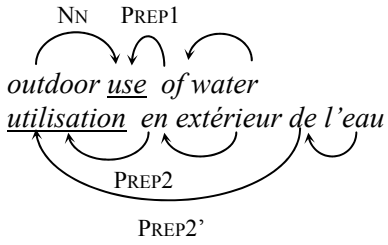
The syntactic propagation may be performed according to two different directions. Indeed, a given word is likely to be both governor and dependent with respect to other words. The first direction consists in starting with dependent anchor words and propagating the alignment link to the governors (DepGov propagation). The DepGov propagation is *a priori* not ambiguous since one dependent is governed at most by one word. Thus, there is just one syntactic relation on which the propagation can be based. The syntactic structures are said to be parallel in English and French provided the two following conditions are met: i) the relation under consideration is identical in both languages and ii) the words involved in the

[3] We use both the French and English versions of the Treetagger. (http://www.ims.uni-stuttgart.de)

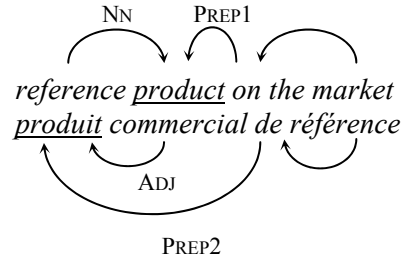
[4] The precision was evaluated manually

syntactic propagation have the same POS. The second direction goes the opposite way: starting with governor anchor words, the alignment link is propagated to the dependents (GovDep propagation). In this case, several relations which may be used to achieve the propagation are available, as it is possible for a governor to have more than one dependent, and so the propagation is potentially ambiguous. The ambiguity is particularly widespread when performing the GovDep propagation from head nouns to their nominal and adjectival dependents. Let us consider the example (1). There is one occurrence of the relation PREP in English and two in French. Thus, it is not possible to determine *a priori* whether to propagate using the relations NN/PREP2, on the one hand, and PREP1/PREP2', on the other hand, or NN/PREP2' and PREP1/PREP2. Moreover, even if there is just one occurrence of the same relation in each language, it does not mean that the propagation is of necessity performed through the same relation, as shown in example (2).

(1)



(2)



In the following sections, we describe precisely the implementation of the two types of propagation defined above in order to align verbs (section 6.2), on the one hand, and nouns and adjectives, on the other hand (section 6.3). To this, we rely on different propagation patterns. Propagation patterns are given in the form CDep-REL-CGov, where CDep is the POS of the dependent, REL is the syntactic relation and CGov, the POS of the governor. The anchor element is underlined and the one aligned by propagation is bolded. For instance, the pattern N-SUJ-**V** corresponds to the propagation going from a noun anchor pair to the verbs through the subject relation.

6.2 Alignment of verbs

Verbs are aligned according to eight propagation patterns, that is to say five for the DepGov propagation and three for the GovDep one.

DEPGOV PROPAGATION TO ALIGN GOVERNOR VERBS. Five propagation patterns are used to align verbs: Adv-MOD-**V** (1), N-SUJ-**V** (2), N-OBJ-**V** (3), N-PREP-**V** (4) and V-PREP-**V** (5).

*(1) The net **is** then **hauled** to the shore.*
*Le filet **est** ensuite **halé** à terre.*
*(2) The fish **are** generally **caught** when they migrate from their feeding areas.*
*Généralement les poissons **sont capturés** quand ils migrent de leur zone d'engraissement.*
*(3) Most of the young shad **reach** the sea.*
*La plupart des alosons **gagne** la mer.*
*(4) The eggs are very small and **fall** to the bottom.*
*Les oeufs de très petite taille **tombent** sur le fond.*
*(5) X is a model which **was designated** to stimulate…*
*X est un modèle qui **a été conçu** pour stimuler…*

GOVDEP PROPAGATION TO ALIGN DEPENDENT VERBS. The alignment links are propagated from the dependents to the verbs using three propagation patterns: **V**-PREP-V (1), **V**-PREP-N (2) and **V**-PREP-Adj (3).

*(1) Ploughing tends to **destroy** the soil microaggregated structure.*
*Le labour tend à **rompre** leur structure microagrégée.*
*(2) The capacity to **colonize** the digestive mucosa…*
*L'aptitude à **coloniser** le tube digestif…*
*(3) An established infection is impossible to **control**.*
*Toute infection en cours est impossible à **maîtriser**.*

	DepGov propagation	GovDep propagation
INRA		
precision (%)	94.1	96.7
JOC		
precision (%)	92.7	97.5

Table 2: Alignment of verbs by means of the DepGov and GovDep propagation

6.3 Alignment of adjectives and nouns

As for verbs, the two types of propagation described in section 6.1 are used to align adjectives and nouns. However, as far as these categories of words are concerned, they can't be treated in a

fully independent way when propagating from head noun anchor words in order to align the dependents. Indeed, the syntactic structure of noun phrases may be different in English and French, since they rely on a different type of composition to produce compounds and on the same one to produce free noun phrases (Chuquet and Paillard, 1989). Then the potential ambiguity arising from the GovDep propagation from head nouns evoked in section 6.1 may be accompanied by variation phenomena affecting the category of the dependents, called transposition (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989). For instance, a noun may be rendered by an adjective, or vice versa: *tax treatment profits* is translated by *traitement fiscal des bénéfices*, so the noun *tax* is in correspondence with the adjective *fiscal*. The syntactic relations used to propagate the alignment links are thus different.

In order to cope with the variation problem, the propagation is performed whether the syntactic relations are identical in both languages or not, and if they are not, whether the categories of the words to be aligned are the same or not. To sum up, adjectives and nouns are aligned *separately* of each other by means of DepGov propagation or GovDep propagation provided that the governor is not a noun. They are *not* treated *separately* when aligning by means of GovDep propagation from head noun anchor pairs.

DEPGOV PROPAGATION TO ALIGN ADJECTIVES. The propagation patterns involved are: Adv-MOD-**Adj** (1), N-PREP-**Adj** (2) and V-PREP-**Adj** (3).

(1) The white cedar exhibits a very ***common*** *physical defect.*
Le Poirier-pays présente un défaut de forme très ***fréquent****.*
(2) The area presently ***devoted*** *to agriculture represents…*
La surface actuellement ***consacrée*** *à l'agriculture représenterait…*
(3) Only fours plots were ***liable*** *to receive this input.*
Seulement quatre parcelles sont ***susceptibles*** *de recevoir ces apports.*

DEPGOV PROPAGATION TO ALIGN NOUNS. Nouns are aligned according to the following propagation patterns: Adj-ADJ-**N** (1), N-NN-**N**/N-PREP-**N** (2), N-PREP-**N** (3) and V-PREP-**N** (4).

(1) Allis shad remain on the continental ***shelf****.*
La grande alose reste sur le ***plateau*** *continental.*
(2) Nature of micropolluant ***carriers****.*
La nature des ***transporteurs*** *des micropolluants.*
(3) The ***bodies*** *of shad are generally fusiform.*
Le ***corps*** *des aloses est généralement fusiforme.*
(4) ***Ability*** *to react to light.*
Capacité *à réagir à la lumière.*

	DepGov propagation	
	Adjectives	Nouns
INRA		
precision (%)	98.7	94.2
JOC		
precision (%)	97.2	93.7

Table 3: Alignment of adjectives and nouns by means of the DepGov propagation

UNAMBIUOUS GOVDEP PROPAGATION TO ALIGN NOUNS. The propagation is not ambiguous when dependent nouns are not governed by a noun. This is the case when considering the following three propagation patterns: **N**-SUJ|OBJ-V (1), **N**-PREP-V (2) and **N**-PREP-Adj (3).

(1) The ***caterpillars*** *can inoculate the* ***fungus****.*
Les ***chenilles*** *peuvent inoculer le* ***champignon****.*
(2) The roots are placed in ***tanks****.*
Les racines sont placées en ***bacs****.*
(3) Botrysis, a fungus responsible for grey ***rot****.*
Botrysis, champignon responsable de la ***pourriture*** *grise.*

POTENTIALLY AMBIGUOUS GOVDEP PROPAGATION TO ALIGN NOUNS AND ADJECTIVES. As we already explained in section 6.1, the propagation is potentially ambiguous when starting with head noun anchor words and trying to align the noun(s) and/or adjective(s) they govern. Considering this potential ambiguity, the algorithm which supports GovDep propagation form head noun anchor words (*n1, n2*) takes into account three situations which are likely to occur :

1. if each of *n1* and *n2* have only one dependent, respectively *reg1* and *reg2,* involving one of the following relations NN, ADJ or PREP; *reg1* and *reg2* are aligned;

   the drained whey
   le lactosérum d'égouttage
   ⇒ (***drained, égouttage***)

2. *n1* has one dependent *reg1* and *n2* several ones $\{reg2_1, reg2_2, \ldots, reg2_n\}$, or vice versa. For each $reg2_i$, check if one of the possible alignments has already been

performed, either by propagation or anchor word spotting. If such an alignment exists, remove the others (*reg1, reg2*$_k$) such as k ≠ i, or vice versa. Otherwise, retain all the alignments (*reg1, reg2*$_i$), or vice versa, without solving the ambiguity;

stimulant substances which are absent from…
substances solubles stimulantes absentes de…
(*stimulant,* {*soluble, stimulant, absent*})
already_aligned(*stimulant, stimulant*) = 1
⇒ (***stimulant***, ***stimulant***)

3. both *n1* and *n2* have several dependents, {*reg1*$_1$, *reg1*$_2$, …, *reg1*$_m$} and {*reg2*$_1$, *reg2*$_2$, …, *reg2*$_n$} respectively. For each *reg1*$_i$ and each *reg2*$_j$, check if one/several alignments have already been performed. If such alignments exist, remove all the alignments (*reg1*$_k$, *reg2*$_l$) such as k ≠ i or l ≠ j. Otherwise, retain all the alignments (*reg1*$_i$, *reg2*$_j$) without solving the ambiguity.

unfair trading practices
pratiques commerciales déloyales
(*unfair,* {*commercial, déloyal*})
(*trading,* {*commercial, déloyal*})
already_aligned(*unfair, déloyal*) = 1
⇒ (***unfair***, ***déloyal***)
⇒ (***trading***, ***commercial***)

a big rectangular net, which is lowered…
un vaste filet rectangulaire immergé…
(*big,* {*vaste, rectangulaire, immergé*})
(*rectangular,* {*vaste, rectangulaire, immergé*})
already_aligned(*rectangular, rectangulaire*) = 1
⇒ (***rectangular, rectangulaire***)
⇒ (*big,* {*vaste, immergé*})

The implemented propagation algorithm has two major advantages: it allows to solve some alignment ambiguities taking advantage of alignments which have been performed previously. This algorithm allows also to cope with the problem of non correspondence between English and French syntactic structures and makes it possible to align words using different syntactic relations in both languages, even though the category of the words under consideration is different.

	GovDep propagation	
	Gov≠Noun	Gov=Noun
INRA		
precision (%)	95.4	97.7
JOC		
precision (%)	95	95.4

Table 4: Alignment of adjectives and nouns by means of the GovDep propagation

6.4 Overall results

Table 5 gives a summary of the results obtained by applying all propagation patterns according to each corpus. It can be seen that the highest accuracy is achieved for the alignments corresponding to anchor pairs validated by the syntactic propagation (AP and PP): 99.7 and 99.8% precision, respectively for INRA and JOC. The rates tend to decrease – respectively 88.5 and 86.1% – as regards alignments established only by means of propagation, referred to as propagated pairs (PP), and is even lower – 76.3% – for the anchor pairs which have not been confirmed by the propagation (AP). Furthermore, the new alignments produced account for less than 20% of overall alignments to approximately 50% for the confirmed ones. Finally, since the method aims at aligning content words, the recall is assessed in relation to their overall occurrences in the corpora.

	Total	AP	AP and PP	PP
INRA				
alignments	50438 (100%)	23646 (47%)	18923 (37%)	7868 (16%)
precision (%)	94.3	76.3	99.7	88.5
recall (%)	58			
JOC				
alignments	71814 (100%)	37118 (52%)	21625 (30%)	13073 (18%)
precision (%)	93.1	94	99.8	86.1
recall (%)	56			

Table 5: overall results of word alignment

7 Discussion

The results achieved by the syntactic propagation method are quite encouraging. They show a high global precision rate – 94.3% for the INRA corpus and 93.1% for the JOC – assessed respectively against a reference list of approximately 8000 and 4600 alignments.

Various reasons make it difficult to compare the results of this experiment with those reported in the literature and presented in section 2. Indeed, each approach has been tested on a different corpus and the results achieved could depend on the type of texts used for evaluation purposes. Moreover, the reference alignment lists, i.e. the gold standards, have probably been established according to different annotation criteria, which could also influence the quality of the results. Finally, each system has been designed, or at least used, to perform a specific task and evaluated in this respect. Daille, Gaussier and Langé (1994), as well as Gaussier (1998) and Hull (2001), were interested in bilingual terminology extraction so that word alignment could not be considered as an end in itself but rather as a basis for term alignment. The system proposed by Wu (2000) aims at bilingual language modelling, word and phrase alignment is incorporated as a subtask. Finally, Arhenberg, Andersson and Merkel (2000) as well as Lin and Cherry (2003) addressed the problem of full word alignment without restricting themselves to content words. Both noticed that the integration of linguistic knowledge, morphological and lexical for the former, syntactic for the latter, improves the alignment quality. However, concerning the approach proposed by Lin and Cherry (2003), it should be pointed out that linguistic knowledge is considered secondary to statistical information. As regards the alignment by syntactic propagation, linguistic knowledge is the kernel of the approach rather than an additional information.

The propagation of alignments links using syntactic relations has proved very efficient when the same propagation pattern is used in both languages, i.e. when the syntactic structures are identical. A high level of precision is also achieved in the case of noun/adjective transpositions, even if the category of the words to be aligned varies. We are actually pursuing the study of non-correspondence between syntactic structures in English and French outlined in (Ozdowska and Bourigault, 2004). The aim is to determine whether there are some regularities in rendering certain English structures into certain French ones or not. If variation across languages is subjected to such regularities, the syntactic propagation could then be extended to the cases of non correspondence.

References

Ahrenberg L., Andersson M. and Merkel M. 2000. A knowledge-lite approach to word alignment, Véronis J. (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 97-138.

Altenberg B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences, Hasselgard and Oksefjell (eds), pp. 249-268.

Bourigault D. and Fabre C. 2000. Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, pp. 131-151, Université Toulouse le Mirail.

Brown P., Della Pietra S. and Mercer R. 1993. *The mathematics of statistical machine translation : parameter estimation*, Computational Linguistics, 19(2), pp. 263-311.

Chuquet H. and Paillard M. 1989. *Approche linguistique des problèmes de traduction anglais/français*, Ophrys.

Daille B., Gaussier E. and Langé J.-M. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology, *Proceedings of the International Conference on Computational Linguistics* (COLING'94) , pp. 515-521.

Debili F. 1997. L'appariement :quels problèmes ?, *Actes des 1ères JST 1997 FRANCIL de l'AUPELF-UREF*, pp. 199-206.

Debili F. and Zribi A. 1996. Les dépendances syntaxiques au service de l'appariement des mots, *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle* (RFIA'96).

Fluhr C., Bisson B. and Elkateb F. 2000. Parallel Text Alignment Using Crosslingual Information Retrieval Techniques, Véronis, J. (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers.

Fox H. J. 2002. Phrasal Cohesion and Statistical Machine Translation, *Proceedings of EMNLP-02*, pp. 304-311.

Frérot C., Bourigault D. and Fabre C. 2003. Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de », in *Traitement Automatique des Langues*, 44-3.

Frérot C., Rigou G. and Lacombe A. 2001. Approche phraséologique d'une extraction automatique de terminologie dans un corpus scientifique bilingue aligné, *Actes des 4èmes rencontres Terminologie et Intelligence Artificielle*, Nancy, pp 180-188.

Gale W. A. and Church K. W. 1991. Identifying Word Correspondences in Parallel Text,

Proceedings of the DARPA Workshop on Speech and Natural Language.

Gaussier E. 1998. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora, *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics* (COLING/ACL'98) , pp. 444-450.

Gaussier E. 2001. General considerations on bilingual terminology extraction, D. Bourigault, Ch. Jacquemin, M.–C. L'Homme (Eds.), *Recent Advances in Computational Terminology*, John Benjamins, pp. 167-183.

Harris B. 1988. Bi-text, A New Concept in Translation Theory, *Language Monthly*, 54, pp.8-10.

Hull D. 2001. Software tools to support the construction of bilingual terminology lexicons, Bourigault, D., Jacquemin, Ch. and L'Homme, M.-C. (Eds.), *Recent Advances in Computational Terminology*, John Benjamins, pp. 225-244.

Lin D. and Cherry C. 2003a. Linguistic Heuristics in Word Alignment, *Proceedings of PACLing 2003*.

Lin D. and Cherry C. 2003b. ProAlign: Shared Task System Description, *Workshop Proceedings on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond (HLT-NAACL 2003)*.

Mihalcea R. and Pedersen T. 2003. An Evaluation Exercise for Word Alignment, *Workshop Proceedings on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond (HLT-NAACL 2003)*, pp. 1-10

Och F. Z. and Ney H., 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1), pp. 19-51.

Ozdowska S. and Bourigault D. 2004. Détection de relations d'appariement bilingue entre termes à partir d'une analyse syntaxique de corpus, *Actes du 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence artificielle*

Véronis J. (Ed). 2000. *Parallel Text Processing : Alignment and Use of Parallel Corpora*, Dordrecht : Kluwer Academic Publishers.

Véronis J. and Langlais P. 2000. Evaluation of parallel text alignment systems. The ARCADE project, Véronis J. (ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 371-388

Vinay J-P. and Darbelnet J. 1958. *Stylistique comparée du français et de l'anglais* , Paris, Didier.

Wu D. 2000. Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars, Véronis, J. (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 139-167.